P19291.P03

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant :A. ITAI

Serial No. :Not Yet Assigned PCT Branch

Filed :Concurrently Herewith PCT/JP98/04457

For :METHOD OF INFERRING THREE-DIMENDSIONAL STRUCTURE OF PROTEIN

CLAIM OF PRIORITY

Commissioner of Patents and Trademarks
Washington, D.C. 20231

Sir:

Applicant hereby claims the right of priority granted pursuant to 35 U.S.C. 119 based upon Japanese Application No. 269611/1997, filed October 2, 1997. The International Bureau already should have sent a certified copy of the Japanese application to the United States designated office. If the certified copy has not arrived, please contact the undersigned.

Respectfully submitted,
A. ITAI

Leslie J. Paperner Reg No. 33,329
Bruce H. Bernstein
Reg. No. 29,027

April 3, 2000
GREENBLUM & BERNSTEIN, P.L.C.
1941 Roland Clarke Place
Reston, VA 20191
(703) 716-1191

PCT/JP98/04457

09/508527

日 本 国 特 許 庁
PATENT OFFICE
JAPANESE GOVERNMENT

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

EU

出願年月日
Date of Application: １９９７年１０月 ２日

出願番号
Application Number: 平成 ９年特許願第２６９６１１号

出願人
Applicant (s): 板井 昭子

PRIORITY DOCUMENT

１９９８年１１月 ６日

特許庁長官
Commissioner,
Patent Office 伴佐山建志

出証番号 出証特平１０－３０８８８８０

【書類名】　特許願

【整理番号】　97171M

【提出日】　平成 9年10月 2日

【あて先】　特許庁長官　殿

【発明の名称】　蛋白質の立体構造の推定方法

【請求項の数】　11

【発明者】

【住所又は居所】　東京都文京区本郷５－１６－６

【氏名】　板井　昭子

【特許出願人】

【識別番号】　592101792

【氏名又は名称】　板井　昭子

【代理人】

【識別番号】　100096219

【弁理士】

【氏名又は名称】　今村　正純

【選任した代理人】

【識別番号】　100092635

【弁理士】

【氏名又は名称】　塩澤　寿夫

【手数料の表示】

【予納台帳番号】　038357

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　9507095

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　蛋白質の立体構造の推定方法

【特許請求の範囲】

【請求項１】　立体構造が既知又は推定可能な参照蛋白質のアミノ酸配列に含まれる各アミノ酸残基の側鎖についての環境情報を含むデータベースを用い、参照蛋白質の各アミノ酸残基の環境情報と質問配列中の各アミノ酸残基の側鎖の疎水性又は親水性の性質とに基づいてマッチングを行い、参照蛋白質の中から質問配列の蛋白質と立体構造の類似性が高い鋳型蛋白質を選択して質問配列の蛋白質のスキャッフォールドを推定する方法。

【請求項２】　参照蛋白質のアミノ酸配列が該参照蛋白質の立体構造に基づいて連続した２以上のアミノ酸残基からなる２以上の部分配列に分割された請求項１に記載の方法。

【請求項３】　参照蛋白質のアミノ酸配列が疎水コアの形成に実質的に関与する１又は２以上のコア部分配列と疎水コアの形成に実質的に関与しない１又は２以上のサブ部分配列とに分割された請求項２に記載の方法。

【請求項４】　参照蛋白質における各アミノ酸残基の側鎖の蛋白質内部への埋没度及び／又は蛋白質表面への露出度の情報と、質問配列の各アミノ酸の疎水性及び／又は親水性の性質とに基づいてマッチングを行う請求項１ないし３のいずれか１項に記載の方法。

【請求項５】　参照蛋白質の１又は２以上のコア部分配列を質問配列上でスライドさせ、該コア部分配列の片端又は両端以外においてはギャップを考慮せずにマッチングを行う請求項１ないし４のいずれか１項に記載の方法。

【請求項６】　ギャップが１又は２以上のアミノ酸残基の削除又は付加である請求項５に記載の方法。

【請求項７】　マッチングが以下の工程：

(a) １又は２以上のコア部分配列を質問配列上でスライドさせながら、必要に応じて該コア部分配列の片端又は両端においてはギャップを考慮してマッチングを行う工程（ただし、２以上のコア部分配列を用いる場合には、該コア部分配列は参照蛋白質のアミノ酸配列での出現順に順番に配置する）；及び

(b) 工程(a) に続いて、必要に応じてギャップを考慮しつつサブ部分配列を質問配列上でスライドさせる工程を含む請求項１ないし６のいずれか１項に記載の方法。

【請求項８】　参照蛋白質のアミノ酸残基の側鎖に関する環境情報と質問配列上の対応アミノ酸残基の疎水パラメータとから計算したスコアに基づいて最適マッチングを選択する、請求項１ないし７のいずれか１項に記載の方法。

【請求項９】　参照蛋白質についての自己マッチングスコアを用いて上記スコアを規格化する工程を含む請求項８に記載の方法。

【請求項１０】　さらに質問配列の蛋白質の立体構造を構築する工程を含む請求項１ないし９のいずれか１項に記載の方法。

【請求項１１】　立体構造が既知又は推定可能な参照蛋白質のアミノ酸配列に含まれる各アミノ酸残基の側鎖についての環境情報を含むデータベースであって、請求項１ないし10のいずれか１項の方法に用いるためのデータベース。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、蛋白質の立体構造の推定方法に関するものである。

【０００２】

【従来の技術】

アミノ酸配列から蛋白質の立体構造を推定することは理論的に不可能とはいえない。しかしながら、現在のところ、配列情報から蛋白質の立体構造を確実に推定する手段は開発されておらず、蛋白質の立体構造を知るための手段は、Ｘ線結晶構造解析やｎｍｒ解析などの実験的方法に限られている。蛋白質の立体構造の情報は、その機能を原子レベルで理解し、またその蛋白質を標的とする医薬の創製やさらに優れた機能をもつ有用な蛋白質の創製に不可欠である。近年、遺伝子情報の解析手段が急速に進歩した結果、実際に蛋白質が単離されないまま配列情報だけが解明される例が急増している。従って、配列情報から蛋白質の立体構造や機能を推定する有効な手段の開発が切望されているのが現状である。

【０００３】

あるアミノ酸配列を有する蛋白質の存在がわかった場合、配列データベースから相同性のある蛋白質を検索するのが一般的である。アミノ酸配列の一致度がある程度よい蛋白質が見つかると、さらにその蛋白質との間で相同性やギャップも考慮したアラインメントを行い、さらに相同性の高いアラインメントの探索が行われる。目的の蛋白質と機能既知の蛋白質との相同性が高い場合には機能がその既知蛋白質に類似しており、一方、立体構造既知の蛋白質との相同性が高い場合には立体構造がその蛋白質に類似しているとの推定が成り立つ。また、相同性が高ければ高いほど機能や立体構造が類似している確率が高く、その推定の信頼性が高いと考えられる。

【０００４】

立体構造既知の蛋白質の配列とある程度（一般に30% 程度）以上の相同性が認められた場合には、その立体構造を鋳型として立体構造を構築するホモロジーモデリング法が行われる。鋳型の立体構造に基づいて対応付けられた残基が鋳型と異なる場合には、側鎖を置換することによって仮想的な立体構造を構築することができる。アラインメント上のギャップは、鋳型立体構造中に対応するアミノ酸がないか、あるいは鋳型の方が余計なアミノ酸をもつことになるので別途処理する必要があるが、ギャップの存在は鋳型に基づくモデリング作業を難しくし、かつ信頼性を損なうので、できるだけギャップの数を少なくするようにギャップにペナルティを与えるアラインメント法が推奨されている。

【０００５】

問題のアミノ酸配列とある程度以上相同性の高いアミノ酸配列を有する立体構造既知の蛋白質が見つからない場合には、ホモロジーモデリングは不可能である。もっとも、蛋白質の結晶構造情報が蓄積されるにつれて、アミノ酸配列の相同性が殆どなく機能も全く異なる複数の蛋白質が、類似の立体構造を持つ例が多数わかってきた。このことは、蛋白質が安定な立体構造をとるための物理化学的要因を考慮すれば、アミノ酸配列の相同性が低い場合にも、立体構造既知の蛋白質群中から鋳型となる立体構造を見出せる可能性があることを示している。

【０００６】

近年、アミノ酸残基毎の疎水性など物理的な性質の一致を考慮したスコアを用

いることにより、アミノ酸配列の相同性が低くても立体構造的に類似性が高いと推定される鋳型蛋白質を立体構造既知の蛋白質群から選ぶ方法が開発された。代表的な方法として、Eisenberg らによる 3D-1D法（R. Luthy, J.U. Bowie and D. Eisenberg, Nature, 356, 83, 1992）がある。この方法は、アミノ酸配列の相同性に加え、立体構造既知の蛋白質について各アミノ酸残基の属する２次構造とその残基位置の環境を表すパラメータと、問題の配列の各アミノ酸残基に対して２次構造毎に与えたパラメータとを用いて、対応づけられたアミノ酸残基間で類似性のスコアを計算する工程を含んでいる。この方法では、蛋白質のペプチド鎖の折り畳み方についての膨大な自由度の問題を既知の結晶構造群を鋳型として用いることで回避しており、評価の要素として疎水性などの物理的パラメータを含めることによって配列の相同性が低い場合にもモデリングが可能である。

【０００７】

しかしながら、立体構造が類似している場合であっても、アミノ酸残基数、２次構造、及びループの長さが同じことは少ないので、アミノ酸配列間の対応づけに基づいた 3D-1D法を実際に適用する場合には数多くの問題が生じる。例えば、アミノ酸配列間の単純なスライド（スレッディング）に加え、どちらかの配列に部分的な配列の欠損など（ギャップ）を考慮して対応付けする必要があるが、ホモロジーモデリングと同様にギャップの導入はモデリングの信頼性を低下させるてしまう。配列の相同性が低いときに、必要にして最小限のギャップを考慮した対応付けをいかにして実現するかが問題である。また、疎水性親水性などのパラメータの他、２０種類のアミノ酸残基に対して２次構造毎のパラメータを用いるなど非常に多数のパラメータに依存しており、パラメータの改良による予測性の向上は期待しにくい。

【０００８】

アミノ酸配列から蛋白質の立体構造を予測する研究の歴史は、配列のどの部分がどのような２次構造をとるかを予測することから始まった。多数の蛋白質の結晶構造情報から統計的に求められたアミノ酸残基毎あるいは連続した数アミノ酸残基の組毎に、α－ヘリックス又はβ－シートになり易さを示すパラメータを用いて、問題のアミノ酸配列から顕著な傾向を示す連続した領域を検出し、それが

どちらの構造をとるかを予測するものである。その代表的なものとして、 Chou と Fasman らによる２次構造予測法（P.Y. Chou, & G. D. Fasman, Adv. enzymol. 47, 45, 1978)がある。しかしながら、このような方法は、２次構造の３次元的集合に関してはなんら情報を与えるものではなく、また予測された２次構造と結晶解析で確認された２次構造の一致が60% 前後であることから、立体構造の推定方法としてはほとんど利用価値がない。

【０００９】

蛋白質の安定な折り畳み構造を、純粋に計算的手法によって先入観を入れずに予測する試みも行われるようになった（いわゆるab initio 予測法）。しかしながら、蛋白質は極めて自由度が大きい分子であり(100残基程度の蛋白質でも考慮すべき自由度のパラメータは400 以上である）、すべての自由度を考慮して可能な構造を十分に探索することは、現在利用可能なコンピュータでは不可能である。また、可能な立体構造の安定性を正しく評価できるほど蛋白質の構造の安定化に関わるファクター（例えば、水の物理化学的性質、疎水相互作用、静電相互作用）についての研究は進んでいないなどの理由から、このような構造予測法の成功は今のところ期待できない。

【００１０】

一方、近年、多数の蛋白質の立体構造が解析されており、その成果はプロテインデータバンクから利用できるようになっている。現在約６,000の蛋白質・核酸の構造が収録されており、機能の異なる独立の蛋白質は400 程度である。これらの蛋白質の立体構造から、配列の相同性もなく進化的にも機能的にも全く関係がないように見える蛋白質が同じ構造モチーフを有する例が多数明らかにされている。

【００１１】

【発明が解決しようとする課題】

本発明の第一の課題は、あるアミノ酸配列を有する蛋白質のアミノ酸配列情報から、その蛋白質がとる可能性の高いスキャッフォールドを推定することにより、立体構造をモデリングするための方法を提供することにある。正しいスキャッフォールドの推定は、立体構造を正しく必要な精度でモデリングするための出発

点となり得る。すなわち、本発明の究極の課題は、アミノ酸配列の情報のみから蛋白質の立体構造を高い信頼性をもって推定する方法を提供することにあり、その手段としてスキャッフォールドを推定する方法を提供することが本発明の具体的課題である。また、本発明の別の課題は、上記の方法に有用なデータベースを提供することにある。

【００１２】

【課題を解決するための手段】

本発明者らは上記の課題を解決すべく鋭意努力した結果、立体構造が既知または推定可能な蛋白質について各アミノ酸残基の側鎖の環境情報を備えたデータベースを作成し、そのデータベースを利用することによって、立体構造未知の蛋白質のアミノ酸配列の情報から、その蛋白質がとる可能性の高いスキャッフォールドを信頼性高く効率的に推定することができる方法を見出した。

【００１３】

すなわち本発明は、立体構造が既知又は推定可能な参照蛋白質のアミノ酸配列に含まれる各アミノ酸残基の側鎖についての環境情報を含むデータベースを用い、参照蛋白質の各アミノ酸残基の環境情報と質問配列中の各アミノ酸残基の側鎖の疎水性又は親水性の性質とに基づいてマッチングを行い、参照蛋白質の中から質問配列の蛋白質と立体構造の類似性が高い鋳型蛋白質を選択して質問配列の蛋白質のスキャッフォールドを推定する方法を提供するものである。スキャッフォールドを推定した後に、鋳型蛋白質と質問配列の最適マッチングに基づいて質問配列に対応した立体構造（３次元座標）を得ることができる。

【００１４】

本発明の好ましい態様によれば、参照蛋白質のアミノ酸配列が該参照蛋白質の立体構造に基づいて連続した２以上のアミノ酸残基からなる２以上の部分配列に分割された上記方法；参照蛋白質のアミノ酸配列が疎水コアの形成に実質的に関与する１又は２以上のコア部分配列と疎水コアの形成に実質的に関与しない１又は２以上のサブ部分配列とに分割された上記方法；参照蛋白質における各アミノ酸残基の側鎖の蛋白質内部への埋没度及び／又は蛋白質表面への露出度の情報と、質問配列の各アミノ酸の疎水性及び／又は親水性の性質とに基づいてマッチン

グを行う上記方法；参照蛋白質の１又は２以上のコア部分配列を質問配列上でスライドさせ、該コア部分配列の片端又は両端以外においてはギャップを考慮せずにマッチングを行う上記方法；並びに、ギャップが１又は２以上のアミノ酸残基の削除又は付加である上記方法が提供される。

【００１５】

本発明のさらに好ましい態様によれば、マッチングが以下の工程：

(a) １又は２以上のコア部分配列を質問配列上でスライドさせながら、必要に応じて該コア部分配列の片端又は両端においてはギャップを考慮してマッチングを行う工程（ただし、２以上のコア部分配列を用いる場合には、該コア部分配列は参照蛋白質のアミノ酸配列での出現順に順番に配置する）；及び

(b) 工程(a) に続いて、必要に応じてギャップを考慮しつつサブ部分配列を質問配列上でスライドさせる工程を含む上記方法；参照蛋白質のアミノ酸残基の側鎖に関する環境情報と質問配列上の対応アミノ酸残基の疎水パラメータとから計算したスコアに基づいて最適マッチングを選択する上記方法；並びに、参照蛋白質についての自己マッチングスコアを用いて上記スコアを規格化する工程を含む上記方法が提供される。

【００１６】

別の観点からは、本発明により、立体構造が既知又は推定可能な１又は２以上の参照蛋白質のアミノ酸残基の側鎖についての環境情報を含むデータベースであって上記の各方法に用いられるためのデータベースが提供される。このデータベースは、通常の記憶媒体、例えば磁気ディスク、光ディスク、CD-ROM、磁気テープなどに格納された形態で流通可能であり、該環境情報と質問配列のアミノ酸残基の性質との一致度をスコアとするマッチングによって、質問配列から構成される蛋白質の安定なスキャッフォールドを推定するために用いることができる。

【００１７】

【発明の実施の形態】

本発明の方法は、質問配列から理論的に可能な主鎖の折り畳み方を網羅・探索して評価するかわりに、参照蛋白質のデータベースから質問配列の蛋白質が立体構造的に類似する鋳型蛋白質を選び、鋳型蛋白質のスキャッフォールドから質問

配列の蛋白質の立体構造を推定することを特徴としている。本発明の方法は、一般的には、ワークステーション、パーソナルコンピューターなどの汎用のコンピューターを用いて高速に行うことができる。

【００１８】

本明細書において用いられる用語は、以下に述べる概念を含めて、最も広義に解釈する必要がある。「立体構造」とは３次元座標で表される蛋白質の構造を意味しており、アミノ酸残基の側鎖など存在する全原子を含む場合もあるが、それらの一部を省略することもある。「モデリング」とは、立体構造が実験的に解明されているか否かに係わらず、ある蛋白質について存在可能性の高い立体構造を構築して原子の３次元座標で表現することをいう。

【００１９】

「２次構造」、「構造モチーフ」、及び「スキャッフォールド」などの用語はBranden 及びTooze らの著書に記載されている（Carl Branden and John Tooze, Introduction to Protein Structure, Garland publishing Inc. New York, 1991: 和訳「タンパク質の構造入門」、教育社、1992）。「構造モチーフ」と「スキャッフォールド」は、ともにペプチド主鎖のみのトポロジーを表す点では共通しているが、「構造モチーフ」が２次構造群の組み合わせとペプチド鎖の流れを平面的かつ模式的に表現するのに対して、「スキャッフォールド」は２次構造群の集合状態を含む蛋白質の３次元構造の骨組を意味する。「立体構造」、「構造モチーフ」、及び「スキャッフォールド」の関係を図１に示す。

【００２０】

立体構造を推定したいアミノ酸配列を「質問配列」と呼び、その質問配列から構成される蛋白質を「質問配列の蛋白質」と呼ぶ。立体構造が既知または推定可能な蛋白質であって本発明のデータベースに含まれる蛋白質を「参照蛋白質」と呼び、参照蛋白質のうち、質問配列とのマッチングスコアがよく、質問配列の蛋白質と立体構造の類似性が高い蛋白質として選別された参照蛋白質を「鋳型蛋白質」という（「鋳型候補蛋白質」と呼ぶ場合もある）。鋳型蛋白質のスキャッフォールドは、質問配列の蛋白質の立体構造を構築する際の鋳型として用いられる。

【００２１】

一般に２以上の配列をアミノ酸残基の一致度や相同性が高くなるように対応づけて並置する操作又は結果は「アラインメント」と呼ばれるが（「一致度」とは配列間で対応する残基間の厳密な一致を意味しており、「相同性」とは同等または類似など曖昧さを含めた一致の割合を意味する）、本発明の方法において「マッチング」（又は「対応付け」）という場合には、アミノ酸残基の一致又は相同性によらず、環境情報と性質の一致で残基を対応づけて並置する操作又は結果を意味している。ある対応付けにおける参照蛋白質の各アミノ酸残基の環境と質問配列のアミノ酸残基の性質との一致度を「マッチングスコア」（又は単に「スコア」）とよぶ。「環境情報」は、主として、参照蛋白質の立体構造における各アミノ酸残基の側鎖の蛋白質表面への露出度や存在環境を示す。アミノ酸配列について用いられる「ギャップ」という用語は、２以上のアミノ酸配列間の対応付けにおいていずれか一方の配列に対応するアミノ酸残基がない部分を指し、一方の配列から見ると１又は２個以上のアミノ酸残基の挿入及び／又は欠失していることを意味する。

【００２２】

本発明の好ましい態様では、このデータベースに参照蛋白質毎に立体構造を反映して分割した２以上の部分配列の情報とアミノ酸残基毎の環境情報とを格納しておく。参照蛋白質の環境情報と質問配列中の対応アミノ酸残基の疎水度パラメータとから計算されるマッチングスコアに基づいて２つの配列をマッチングする。疎水度パラメータは２０種のアミノ酸残基に対して予め数値化しておくことが望ましい。

【００２３】

部分配列のうち、疎水コアの形成に関与するコア部分配列についてはギャップを入れずに質問配列上をスライドさせ、両端においてのみアミノ酸残基の増減（隣接するサブ部分配列の末端から１又は２個以上のアミノ酸残基をとって該コア部分配列に組み込むか、又はその逆の操作を意味する）を考慮してマッチングを行い、スコアのよい鋳型候補蛋白質を選別する。次に、疎水コアの安定化に関与しないサブ部分配列について必要に応じてギャップを考慮したマッチングを行い

、鋳型候補の数を絞る。最終的な鋳型蛋白質の選択は、各蛋白質の最適マッチングのスコアを自己マッチングスコアで規格化して比べることにより行うことができる。

【００２４】

本発明の方法は、マッチングスコアが高いほど２つの配列間でスキャッフォールドや立体構造の類似性が高いとの仮定に基づいており、配列の相同性が殆どない配列間の残基のマッチングを蛋白質の折り畳み原理に従って行うことによって、参照蛋白質からの鋳型蛋白質の適切な選択を可能にしている点に特徴がある。ある。本発明の好ましい態様では、(1) データベースの作成；(2) 部分配列を用いたマッチング；(3) マッチングスコアの計算；(4) 参照蛋白質から鋳型候補蛋白質の選択；(5) 鋳型蛋白質の選択を主要な要素としている。

【００２５】

一般的な水溶性の単一サブユニットの蛋白質を構成するペプチド鎖については、アミノ酸残基の疎水性側鎖ができるだけ分子内部に埋もれて露出せず、一方、親水性側鎖はできるだけ蛋白質分子表面に露出するのが自由エネルギー的に有利である。もっとも、細胞膜に相互作用する蛋白質や複数のサブユニットからなる蛋白質など存在環境が異なる蛋白質では、個々の蛋白質構造やサブユニット構造だけをみると、疎水性側鎖が蛋白質表面に露出していることもある。本発明の方法では、このような蛋白質の立体構造の多様性も考慮されており、個々のアミノ酸残基の存在環境を反映したスコア付けができるようになっている。アミノ酸残基毎の特定２次構造のとり易さの概念（例えば P.Y. Chou, & G. D. Fasman, Adv. enzymol. 47, 45, 1978）は基本的には用いないが、必要に応じてそれらの概念を加味したスコアを用いることができる。

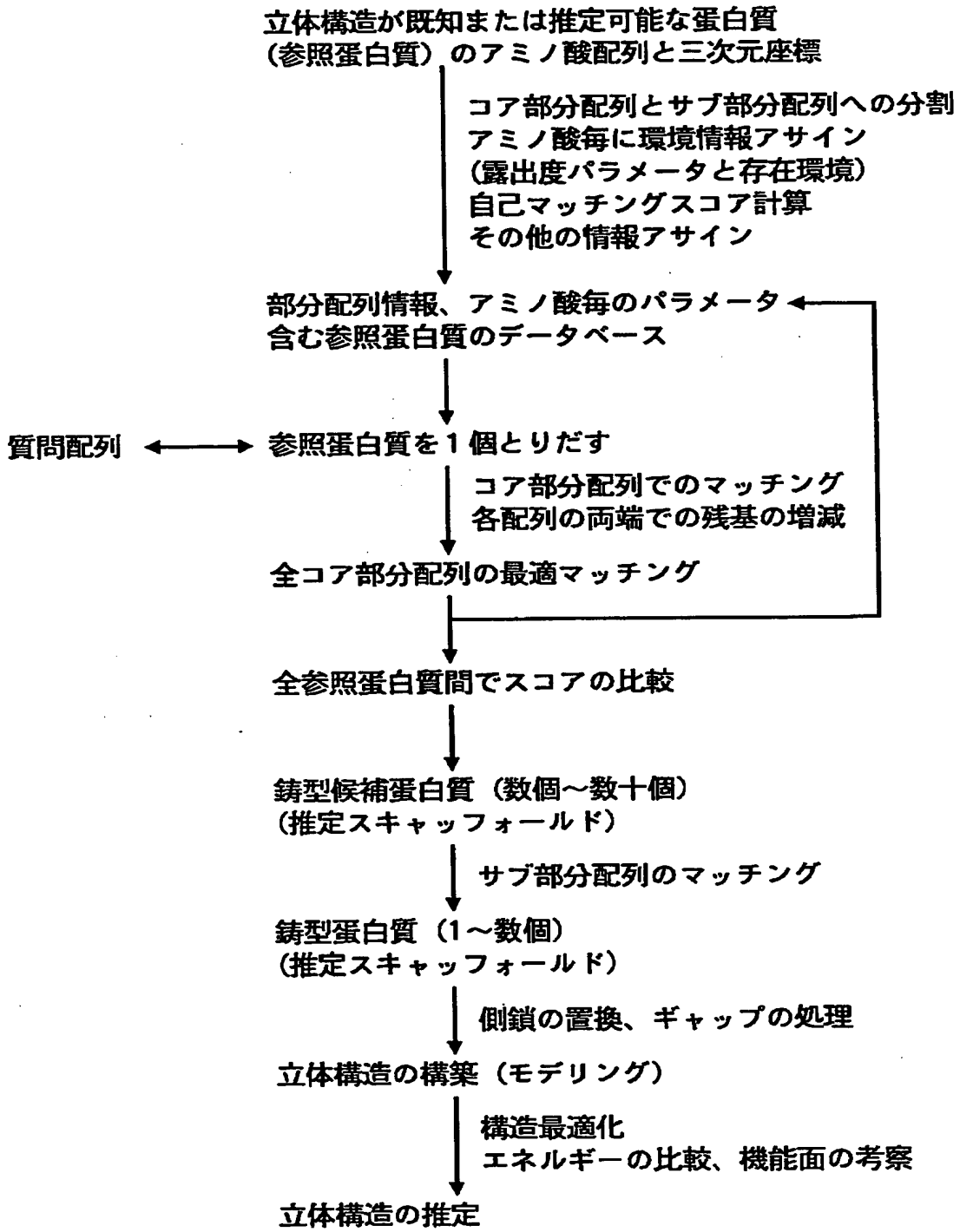
【００２６】

【実施例】

以下、本発明の方法の好ましい態様をスキームで示し、このスキームに従って本発明の方法をより具体的に説明するが、本発明の方法はこのスキームの細部又は以下の説明の細部に限定されることはない。

【００２７】

【化１】


立体構造が既知または推定可能な蛋白質
(参照蛋白質) のアミノ酸配列と三次元座標
コア部分配列とサブ部分配列への分割
アミノ酸毎に環境情報アサイン
(露出度パラメータと存在環境)
自己マッチングスコア計算
その他の情報アサイン
部分配列情報、アミノ酸毎のパラメータ
含む参照蛋白質のデータベース
質問配列
参照蛋白質を１個とりだす
コア部分配列でのマッチング
各配列の両端での残基の増減
全コア部分配列の最適マッチング
全参照蛋白質間でスコアの比較
鋳型候補蛋白質（数個～数十個）
(推定スキャッフォールド)
サブ部分配列のマッチング
鋳型蛋白質（1～数個）
(推定スキャッフォールド)
側鎖の置換、ギャップの処理
立体構造の構築（モデリング）
構造最適化
エネルギーの比較、機能面の考察
立体構造の推定

【００２８】

(A) データベースの作成

立体構造が既知または推定可能な参照蛋白質について、立体構造に関わる情報として、各アミノ酸残基の環境情報及び部分配列の情報を含むデータベースを作成しておく。データベースに収納する蛋白質としては、立体構造の情報が既知または推定可能な蛋白質であれば、すべてエントリーとすることができる。同時に構造決定された蛋白質中に複数のサブユニットが含まれるときは、独立のエントリーとしておくことが好ましい。ペプチド鎖で繋がった複数ドメインからなる構造については、全体構造とともに各ドメインも独立エントリーとしておくことが好ましい。

【００２９】

参照蛋白質毎に含める情報は以下の通りである。

(1) 一般的な情報

(a) 蛋白質名（蛋白質コード）、サブタイプ、アミノ酸数、アミノ酸配列、ドメイン、サブユニットなどに関する情報。

(b) 立体構造に関する情報として、立体構造の決定（又は推定）方法（結晶解析法、ｎｍｒ法、又はモデリング法のいずれにより立体構造を決定したか）、ＰＤＢコード、モデリング法による場合には鋳型として利用した蛋白質分子名、結晶解析の場合には共結晶化された分子がある場合にはその化学名などの情報のうち適宜のものを含めておく。

(c) 生物学的意義に関する情報として、生物学的機能、生物種、存在する組織・臓器、エフェクター分子などの情報を含めておく。

【００３０】

(2) 部分配列情報

蛋白質の立体構造における特徴に従って、配列を分割して２以上の部分配列とし、各部分配列について以下の情報をデータベースに含めるのが望ましい。

(a) Ｎ末端から何番目の部分配列か

(b) コア部分配列かサブ部分配列かのフラグ

(c) 始点と終点（Ｎ末端からのアミノ酸配列番号で）、配列の長さ、その距離及びベクトル、サブ部分配列ならば最短の残基数

(d) 部分配列間の距離及びベクトル

(e) β－シート形成の水素結合をする相手の部分配列番号、および逆平行または平行β－シートなどの区別、あるいは一定の距離内にある部分配列番号など

(f) 自己マッチングスコア（後述）

【００３１】

部分配列への分割の基準は特に限定されないが、α－ヘリックス構造やβ－ストランド構造などの２次構造を形成し実質的に疎水コアの形成に関わる部分をコア部分配列とし、それ以外の部分をサブ部分配列とするのができる。各部分配列は、例えば、７残基以上を含む連続した配列とすることが望ましい。β－ターン構造については、はじめからコアまたはサブ部分配列に加えても構わないし、別のフラグによって区別し、検索時にコアかサブかの判断を加えてもよい。分割の作業は、コンピュータグラフィックス画面上で対話的に個々の蛋白質について行ってもよいし、分割の基準を定めたプログラムを作成して自動的に行ってもよい。その判断基準として、例えば、隣り合う４つのCアルファ原子のなす擬ねじれ角などの数値を基準に用いれば、自動的に部分配列に分割できる。

【００３２】

(3) アミノ酸残基毎の環境情報

(a) 露出度パラメータ

参照蛋白質の全てについて、立体構造に基づいて残基毎の蛋白質表面への側鎖の露出度及び蛋白質内部への側鎖の埋没度を計算し、その計算値に基づいて露出度パラメータを割り振る。本明細書において用いられる「露出度パラメータ」という用語は、各残基の側鎖がどの程度分子表面に露出しているか、又は埋没しているかを示す数値を意味している。露出度パラメータを定義する方法は特に限定されず、いかなる方法を採用してもよいが、露出度が高いものはマイナスの値、露出度が低いものはプラスの値をとるように設定するのが望ましい。例えば、立体構造において各アミノ酸側鎖の溶媒への接触可能表面と蛋白質原子への接触表面積を計算して、その差から露出度パラメータを算出して利用することができる。また、例えば、全分子表面に対する溶媒接触表面の割合を基準として定めることもできる。

【００３３】

(b) 存在環境フラグの設定

細胞膜に相互作用する蛋白質では膜と相互作用する部分の表面に疎水性アミノ酸残基の側鎖が露出しており、また、サブユニットやドメイン構造が集合して安定化する接触面を有する蛋白質では、接触面に疎水性アミノ酸残基の側鎖が露出している場合があり、これらの蛋白質は単独で存在する水溶性蛋白質とは異なる性質を有している。このような蛋白質については、一般的には、先述の折り畳み原理はそのままでは適用できない。データベース作成時に、例えば、次のような存在環境フラグを露出度パラメータとは別に各アミノ酸残基に与えておくことができる。

【００３４】

例えば、蛋白質の由来や機能に関する実験結果および立体構造から、各アミノ酸の存在環境が以下のいずれに相当するかを推定して環境フラグを与えておき、マッチングやスコアの計算の際に考慮することが可能である。

０：不明（未定義または定義できず）

１：分子内部（蛋白質内、サブユニット内、ドメイン内の接触）

２：分子内孔（リガンド結合部位）

３：分子表面（水環境と接触）

４：分子表面（別蛋白質、別サブユニット、別ドメインと接触）

５：分子表面（膜と接触）

【００３５】

さらに、立体構造形成に特殊な影響を与えるアミノ酸残基については、特殊残基であることを示すフラグを与えておき、マッチングやスコアの計算の際に考慮することができる。例えば、Ｓ－Ｓ結合しているシステイン残基やプロリンのように主鎖に水素結合性官能基が不足しているアミノ酸残基、または側鎖原子間に強い親水性相互作用を形成する可能性のある残基などがに適用することができる。

【００３６】

(B) 質問配列と参照蛋白質の配列間のマッチング

一方の配列を他方の配列上スライドさせて最適のマッチングを効率よく探すために、部分配列の概念を利用することができる。そのために、上記データベースは、各参照蛋白質について立体構造から得られる部分配列の情報とアミノ酸残基ごとの環境情報がアミノ酸配列順に収めてある。一方、質問配列が有している情報はアミノ酸配列の情報だけであり、アミノ酸残基毎の疎水パラメータの表からとった値を当てはめてスコアの計算に用いる。データベースから参照蛋白質を1つずつ取り出し、アミノ酸配列中に出現する順に並べた部分配列群を質問配列上でスライドさせて、部分配列群と質問配列との間でマッチングスコアの最もよいマッチングを探す。

【００３７】

(1) 部分配列を用いたマッチング

部分配列を用い、かつコア部分配列ではギャップを考慮しないでアミノ酸残基単位でマッチングを行うことにより、最適マッチングの配列間対応付けを高速に検索することができる。一般に、進化の過程では残基の置換とともに挿入や欠失が起きることが多く、その考慮は配列のマッチングに不可欠である（「従来の技術」の欄を参照）。しかしながら、一般的には、ギャップが入るのはサブ部分配列であることが多い。これは、疎水コアの安定化に関与するコア部分配列でその両端以外で挿入や欠失が起きると、その安定スキャッフォールド自体が損なわれ、蛋白質の立体構造が大きく変化してしまうからである。

【００３８】

そこで、本発明の方法の好ましい態様では、コア部分配列とサブ部分配列とに分けて２段階のマッチングを行う。第一段階のコア部分配列を用いたマッチングにおいては、まずギャップを考慮せずにに１又は２以上のコア部分配列を質問配列上でスライドさせ、該コア部分配列の両端においてのみアミノ酸残基の増減を考慮したマッチングを行って最適マッチングを探索する。

【００３９】

各コア部分配列について質問配列の上をスライドさせながらマッチングスコアを計算して保存する。全部のコア部分配列について同様に計算した後、全体として最適なマッチングを決定する。２以上のコア部分配列を用いる場合には、質問

配列上に２以上のコア部分配列を参照蛋白質のアミノ酸配列中での出現順序に従って重なり合わないように配置し、コア部分配列間には４個程度以上のアミノ酸残基の存在を仮定して（立体的に結合可能なアミノ酸残基数で隣のコア部分配列と順番に繋がるという条件、例えばβーターンなどに要するアミノ酸残基数）、それぞれのコア部分配列の順序を変えずに質問配列上をスライドさせ、最も総スコアのよいマッチングを選択する。この際、各コア部分配列のマッチングスコアが最大である必要はない。この第一段階で、高いスコアを与える鋳型蛋白質を数個から数十個選択し、スキャッフォールド候補として第二段階に進む。

【００４０】

(C) マッチングスコア

マッチングスコアは参照蛋白質の環境情報の露出度パラメータEP(i) と質問配列の残基の疎水パラメータHB(j) を用いて計算する。i は参照蛋白質のアミノ酸配列中のアミノ酸残基番号であり、j はそれと対応づけられた質問配列の残基番号を指す。マッチングスコアの計算式は、参照蛋白質の分子内部に埋没した側鎖環境に質問配列の疎水性の強いアミノ酸残基が対応し、分子表面に露出した側鎖環境に質問配列の親水性の強いアミノ酸残基が対応すると高いスコアが得られるような計算式であれば、いかなるものを利用してもよい。マッチングスコアは例えば、最も単純には次式によって計算することができる。

【００４１】

【数１】

残基毎のマッチングスコア＝EP(i) ×HB(j)

部分配列のマッチングスコア＝配列に含まれる残基のマッチングスコアの和

全配列のマッチングスコア＝全部分配列のマッチングスコアの和

【００４２】

(1) 疎水パラメータ

２０種のアミノ酸残基のそれぞれに疎水性又は親水性の性質に関連した疎水パラメータを与えておく。疎水パラメータの決定方法は特に限定されず、いかなる基準による値を用いてもよい。例えば、文献に記載されたアミノ酸毎の疎水性値を用いてもよく、又は適宜の方法により独自の基準で算出したものを用いてもよい

。また、あるアミノ酸について、結晶解析された蛋白質中の全出現回数に対して蛋白質分子内部に埋没された残基の比率を統計的に求めておき、その比率を該アミノ酸の疎水パラメータとして利用してもよい。

【００４３】

また、例えば、個々のアミノ酸残基に別々の値を与えてもよいが、次のように段階化した疎水パラメータを与えることもできる。

【表１】

２：強い疎水性（イソロイシン、バリン、ロイシン、フェニルアラニン）

１：弱い疎水性（アラニン、メチオニン、シスチン、チロシン）

０：ほぼ中性　（グリシン、プロリン、リジン、アルギニン）

－１：弱い親水性（スレオニン、ヒスチジン）

－２：強い親水性（セリン、アスパラギン、アスパラギン酸、グルタミン、グルタミン酸）

【００４４】

(2) 自己マッチングスコア

アミノ酸数及びアミノ酸組成の異なる蛋白質間で質問配列へのマッチングの良さを比較するためにはスコアの規格化をしておくことが望ましい。そのために、各参照蛋白質について、それ自体のアミノ酸配列の露出度パラメータEP(i) とデータベースに用意された環境情報の疎水パラメータHB(i) とからマッチングスコアを予め計算してデータベースに保存しておく。例えば、次式のように計算するとよい。

【００４５】

【数２】

自己マッチングスコア＝Σ(EP(i)×HB(i))

【００４６】

全部分配列の質問配列へのマッチングが終了して最適マッチングが得られた後、得られたマッチングスコアに自己マッチングスコアを乗じて規格化することができる。全参照蛋白質について規格化された最適マッチングスコア同士を比較して最適な鋳型候補蛋白質を選択することができる。自己マッチングスコア及びマ

ッチングスコアは、アミノ酸残基数が多いほど大きな値を取りやすい。

【００４７】

(D) 鋳型候補蛋白質の選択

鋳型候補蛋白質の選択の手順は概ね以下の通りである。

(a) データベースから参照蛋白質を１個づつ取り出し、質問配列に対しマッチングを行う；

(b) コア部分配列につきギャップを考慮せずに質問配列上をスライドしマッチングスコアを算出する；

(c) 必要に応じて、各コア部分配列のＮ末端またはＣ末端の残基を増減しながらマッチングし、最適マッチングを得る；

(d) 参照蛋白質全部について工程(a) ～(c) を行い、最適マッチングとマッチングスコアを得る；

(e) 規格化したスコアにより参照蛋白質からスコアのよいものを鋳型候補蛋白質とする（この段階で、それらの構造は質問配列のスキャッフォールドの候補とすることができる）；

(f) コア部分配列間をつなぐサブ部分配列のマッチングを行う。質問配列の対応する配列部分との配列長の違いやギャップの存在を考慮し、最適のマッチングとマッチングスコアを得る；及び

(g) 規格化したスコアによって鋳型蛋白質を選択する。

【００４８】

(E) 立体構造の構築

質問配列の蛋白質の立体構造は、鋳型蛋白質のスキャッフォールドの構造、及び該参照蛋白質と質問配列との最適マッチングの結果に基づいて、アミノ酸残基の側鎖の置換を行うことにより構築することができ、質問配列に対応した３次元座標を得ることができる。優劣をつけ難い２以上の鋳型蛋白質がある場合には、すべてについて立体構造を構築することが望ましい。サブ部分配列の長さが鋳型候補と異なる場合には、結晶構造に現れたループ構造を集めたデータベース等を用いて、該サブ部分配列の適切なトポロジーを決定することができる。鋳型のスキャッフォールドが修正される部分については、マッチングスコアと同様に残基

の性質と露出度を考慮しつつトポロジーを決定することができる。重大な立体障害や立体構造を不安定化する分子内部の隙間などの有無を確認し、また構造最適化計算や分子動力学計算により構造の微調整を行ったのちに立体構造の安定性を比較する。

【００４９】

最後に、全ての鋳型に基づいて構築された立体構造をエネルギーとマッチングスコアにより順位付けするが、該質問配列の蛋白質の機能が既知の場合には、該鋳型蛋白質に知られた機能との対応、その機能の発現に関与するものと推定されるアミノ酸残基の立体構造上の位置の妥当性、さらにはアミノ酸変異による機能への影響などの情報などを鋳型蛋白質の選別に利用できる。

【００５０】

【発明の効果】

本発明の方法によれば、立体構造が既知または推定可能な蛋白質のアミノ酸配列データベースに基づいて、アミノ酸配列の情報のみからそのアミノ酸配列により構築される蛋白質の立体構造に関する情報を信頼性高く効率的に入手することができる。

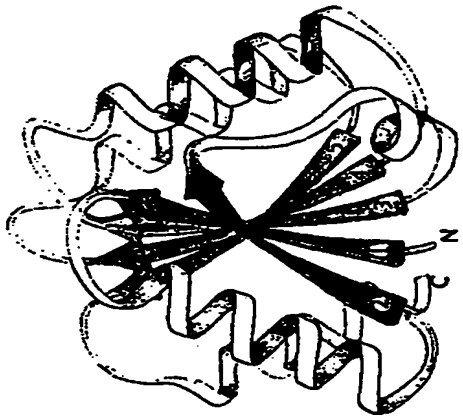
【図面の簡単な説明】

【図１】　「立体構造」、「構造モチーフ」、及び「スキャッフォールド」の関係を示した図である。図中、(a) は立体構造を示し、(b) は構造モチーフを示し、(c) はスキャッフォールドを示す。

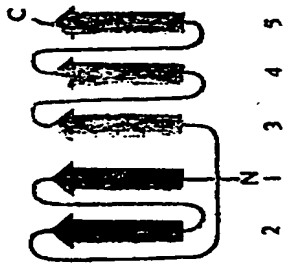
【書類名】　図面

【図１】

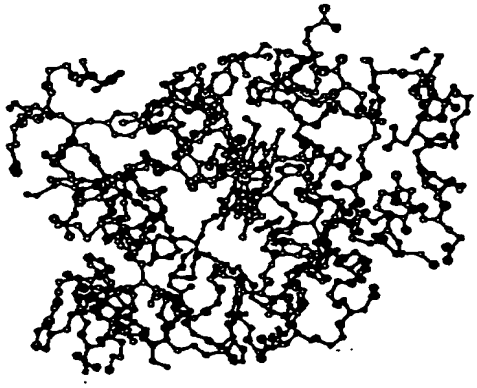
(c)



(b)



(a)



【書類名】　　要約書

【要約】

【解決手段】　立体構造が既知又は推定可能な参照蛋白質のアミノ酸配列に含まれる各アミノ酸残基の側鎖についての環境情報を含むデータベースを用い、参照蛋白質の各アミノ酸残基の環境情報と質問配列中の各アミノ酸残基の側鎖の疎水性又は親水性の性質とに基づいてマッチングを行い、参照蛋白質の中から質問配列の蛋白質と立体構造の類似性が高い鋳型蛋白質を選択して質問配列の蛋白質のスキャッフォールドを推定する方法。

【効果】　アミノ酸配列の情報のみからそのアミノ酸配列により構築される蛋白質の立体構造に関する情報を信頼性高く効率的に入手することができる。

【選択図】　なし

【書類名】　職権訂正データ
【訂正書類】　特許願

＜認定情報・付加情報＞
【特許出願人】
【識別番号】　592101792
【住所又は居所】　東京都文京区本郷５－１６－６
【氏名又は名称】　板井　昭子
【代理人】　申請人
【識別番号】　100096219
【住所又は居所】　東京都中央区八重洲１丁目８番１２号　藤和八重洲一丁目ビル７階
【氏名又は名称】　今村　正純
【選任した代理人】
【識別番号】　100092635
【住所又は居所】　東京都中央区八重洲１丁目８－１２　藤和八重洲一丁目ビル７Ｆ
【氏名又は名称】　塩澤　寿夫

出　願　人　履　歴　情　報

識別番号　　〔592101792〕

1．変更年月日　　1992年　3月27日
［変更理由］　新規登録
住　所　東京都文京区本郷5－16－6
氏　名　板井　昭子